Correction du TD sur la régression logistique

1. Les observations sont $(x_i, y_i)_{i\in\{1,\cdots,900\}}$, où
   — $x_i \in \{0,1,2\}$ est le nombre de paquets consommés par jour par le i-ème individu,
   — $y_i \in \{0,1\}$, avec $y_i = 1$ si le i-ème individu souffre de problèmes de circulation.
   On suppose que $(x_i, y_i)_{i\in\{1,\cdots,900\}}$ est une réalisation d'un échantillon du couple de variables (X, Y), et on veut tester (H_0) : "X et Y sont indépendantes" contre (H_1) : "X et Y sont dépendantes". Comme X et Y sont deux variables qualitatives, on fait un test du χ^2 d'indépendance. Ce test est basé sur la statistique

$$T = \sum_{i\in\{0,1,2\}, j\in\{0,1\}} \frac{\left(N_{ij} - \frac{N_{i\bullet}N_{\bullet j}}{n}\right)^2}{\frac{N_{i\bullet}N_{\bullet j}}{n}},$$

   où pour $i \in \{0,1,2\}$, et $j \in \{0,1\}$,
   — N_{ij} est le nombre d'individus fumant i paquets, pour lesquels $y = j$;
   — $N_{i\bullet}$ est le nombre d'individus fumant i paquets ;
   — $N_{\bullet 0}$ (respectivement $N_{\bullet 1}$) est le nombre d'individus sans (respectivement avec des) problèmes de circulation.
   Lorsque n est grand, la loi de T sous (H_0) est approximativement une loi du χ^2 à $(3-1)(2-1) = 2$ degrés de liberté. La P-valeur du test est donc donnée par $\mathbb{P}[Z \geq T_{obs}]$, où $Z \sim \chi^2_2$.
   Le calcul de T_{obs} donne

$$T_{obs} = \frac{\left(40 - \frac{310(300)}{900}\right)^2}{\frac{310(300)}{900}} + \frac{\left(70 - \frac{310(300)}{900}\right)^2}{\frac{310(300)}{900}} + \frac{\left(200 - \frac{310(300)}{900}\right)^2}{\frac{310(300)}{900}} + \frac{\left(260 - \frac{590(300)}{900}\right)^2}{\frac{590(300)}{900}}$$
$$+ \frac{\left(230 - \frac{590(300)}{900}\right)^2}{\frac{590(300)}{900}} + \frac{\left(100 - \frac{590(300)}{900}\right)^2}{\frac{590(300)}{900}} = 213.55\,.$$

   et on a $P_{val} \leq 5\%$. On conclut donc qu'il y a dépendance entre consommation de tabac et présence de problèmes de circulation avec 5% de chance de se tromper.

2. On choisit comme profil de référence le profil "non fumeur". La dimension explicative x : "nombre de paquets par jours" sera donc codée par 2 variables indicatrices :
   — $x^1 \in \{0,1\}$, avec $x^1_i = 1$ si le i-ème individu fume 1 paquet par jour, et $x^1_i = 0$ sinon ;
   — $x^2 \in \{0,1\}$, avec $x^2_i = 1$ si le i-ème individu fume plus de 2 paquets par jour, et $x^2_i = 0$ sinon.
   Un individu malade et non fumeur sera alors codé par $(x^1 = 0, x^2 = 0, y = 1)$. Un individu sain et fumant plus de 2 paquets par jour sera quant à lui codé par $(x^1 = 0, x^2 = 1, y = 0)$.

3. Pour $x = (x^1, x^2) \in \{0,1\}^2$, le modèle de régression logistique postule que

$$\mathbb{P}(Y = 1 | X = x) = \frac{\exp(\beta_0 + \beta_1 x^1 + \beta_2 x^2)}{1 + \exp(\beta_0 + \beta_1 x^1 + \beta_2 x^2)}\,.$$

Il y a 3 valeurs possibles pour x correspondant aux trois modalités de la consommation de tabac. Ces trois valeurs possibles sont $x = (0,0)$ correspondant à la classe C_0, $x = (1,0)$ correspondant à la classe C_1, $x = (0,1)$ correspondant à la classe C_2. $x = (1,1)$ est impossible car il faudrait à la fois fumer un paquet par jour, et plus de deux.

La classe C_0 est codée par $x^1 = x^2 = 0$, ce qui donne $\pi_0 = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$.

La classe C_1 est codée par $x^1 = 1, x^2 = 0$, soit $\pi_1 = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$.

La classe C_2 est codée par $x^1 = 0, x^2 = 1$, soit $\pi_2 = \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)}$.

4. On a $\text{cote}(C_0) = \text{cote}(x = (0,0)) = \frac{\pi_0}{1-\pi_0} = \frac{\exp(\beta_0)/(1+\exp(\beta_0))}{1/(1+\exp(\beta_0))} = \exp(\beta_0)$, ce qui donne

$$\beta_0 = \log\left(\frac{\pi_0}{1-\pi_0}\right).$$

De la même façon, $\text{cote}(C_1) = \exp(\beta_0 + \beta_1)$ et $\text{cote}(C_2) = \exp(\beta_0 + \beta_2)$. En faisant les rapports de cote, on obtient

$$OR(C_1) = OR(x = (1,0)) = \frac{\text{cote}(x = (1,0))}{\text{cote}(x = (0,0))} = \frac{\text{cote}(C_1)}{\text{cote}(C_0)} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1).$$

On a donc

$$\beta_1 = \log\left(\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}\right).$$

De la même façon, on a $OR(C_2) = \exp(\beta_2)$, soit

$$\beta_2 = \log\left(\frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_0}{1-\pi_0}}\right).$$

5. Les probabilités empiriques d'être malade pour chaque classe sont données par

$$\bar{\pi}_i = \frac{\text{nombre de malades dans } C_i}{\text{effectif de } C_i} = \frac{N_{i1}}{N_{i\bullet}}.$$

On a donc :

$$\bar{\pi}_0 = 40/300 = 2/15\,;\ \bar{\pi}_1 = 70/300 = 7/30\,;\ \bar{\pi}_2 = 200/300 = 2/3\,.$$

6. Les cotes empiriques pour chaque classe sont :
   — $\overline{\text{cote}}(C_0) = \frac{\bar{\pi}_0}{1-\bar{\pi}_0} = 2/13$.
   — $\overline{\text{cote}}(C_1) = \frac{\bar{\pi}_1}{1-\bar{\pi}_1} = 7/23$.
   — $\overline{\text{cote}}(C_2) = \frac{\bar{\pi}_2}{1-\bar{\pi}_2} = 2$.

   Les rapports de cote empiriques sont
   — $\overline{\text{OR}}(C_1) = \frac{\overline{\text{cote}}(C_1)}{\overline{\text{cote}}(C_0)} = \frac{7/23}{2/13} = 91/46 = 1.97$.
   — $\overline{\text{OR}}(C_2) = \frac{\overline{\text{cote}}(C_2)}{\overline{\text{cote}}(C_0)} = \frac{2}{2/13} = 13$.

7. $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ est la valeur de β qui maximise la vraisemblance donnée par

$$V(\beta) = \prod_{i=1}^{n} \mathbb{P}\left[Y = y_i | X = x\right] = \pi_0^{N_{01}}(1-\pi_0)^{N_{00}}\pi_1^{N_{11}}(1-\pi_1)^{N_{10}}\pi_2^{N_{21}}(1-\pi_2)^{N_{20}}.$$

Il revient au même de maximiser la log-vraisemblance

$$\begin{aligned} L(\beta) = \log V(\beta) &= N_{01}\log(\pi_0) + N_{00}\log(1-\pi_0) + N_{11}\log(\pi_1) \\ &\quad + N_{10}\log(1-\pi_1) + N_{21}\log(\pi_2) + N_{20}\log(1-\pi_2) \end{aligned}$$

avec

$$(\pi_0, \pi_1, \pi_2) = (F(\beta_0), F(\beta_0+\beta_1), F(\beta_0+\beta_2)), \quad \text{où } F(x) = e^x/(1+e^x). \qquad (1)$$

Comme la relation (1) liant π et β est bijective, on peut commencer par maximiser L vue comme une fonction de (π_0, π_1, π_2), puis inverser la relation (1) pour obtenir $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$. Calculons le gradient de L en (π_0, π_1, π_2) :

$$\frac{\partial L}{\partial \pi_i} = \frac{N_{i1}}{\pi_i} - \frac{N_{i0}}{1-\pi_i}.$$

Le maximum de L est donc atteint en $(\hat{\pi}_0, \hat{\pi}_1, \hat{\pi}_2)$ solution de

$$\frac{N_{i1}}{\hat{\pi}_i} - \frac{N_{i0}}{1-\hat{\pi}_i} = 0 \Leftrightarrow \hat{\pi}_i = \frac{N_{i1}}{N_{i\bullet}} = \bar{\pi}_i.$$

On en déduit que

$$\hat{\beta}_0 = \log\left(\frac{\bar{\pi}_0}{1-\bar{\pi}_0}\right), \ \hat{\beta}_1 = \log\left(\frac{\frac{\bar{\pi}_1}{1-\bar{\pi}_1}}{\frac{\bar{\pi}_0}{1-\bar{\pi}_0}}\right), \ \hat{\beta}_2 = \log\left(\frac{\frac{\bar{\pi}_2}{1-\bar{\pi}_2}}{\frac{\bar{\pi}_0}{1-\bar{\pi}_0}}\right).$$

Sur notre échantillon de données, on obtient donc

$$\hat{\beta}_0(\omega) = \log(2/13) = -1.87, \ \hat{\beta}_1(\omega) = \log(91/46) = 0.68, \ \hat{\beta}_2(\omega) = \log(13) = 2.56.$$

8. Lorsque $\beta_0 = 0$, $\pi_0 = 1/2$ et on a autant de chances d'être malade que ne de pas l'être quand on est non fumeur. Lorsque β_0 augmente, la cote de la classe C_0 augmente et la part des malades augmente dans la classe C_0. Lorsque β_0 diminue, la cote de la classe C_0 diminue et la part des malades diminue dans la classe C_0.
   Lorsque $\beta_1 = 0$, la cote de la classe C_1 est la même que celle de la classe de référence C_0. Autrement dit, la part de malades est identique dans la classe des non fumeurs et dans celle de ceux qui fument 1 paquet par jour. Lorsque β_1 augmente, la part des malades devient plus importante dans la classe des fumeurs d'un paquet par jour que dans celle des non fumeurs. Une autre façon de voir les choses est de dire que lorsque $\beta_1 = 0$, on a $\pi_1 = \pi_0$ et la probabilité d'être malade est la même qu'on ne fume pas, ou qu'on fume un paquet par jour. Lorsque β_1 augmente dans les positifs, π_1 devient plus grand que π_0 et le probabilité d'être malade est plus importante quand on fume un paquet, que quand on ne fume pas. Si β_1 diminue (dans les négatifs) π_0 devient plus grand que π_1, et on a donc plus de chances d'être malade quand on est non fumeur que lorsqu'on fume un paquet.
9. Notons D_T (respectivement $D_\emptyset$) la déviance du modèle total (respectivement vide). Par définition, pour un modèle M, $D_M = -2L_M(\hat{\beta}_M)$, où L_M est la logvraisemblance associée au modèle M, et $\hat{\beta}_M$ est l'estimateur du maximum de vraisemblance associé. Pour le modèle total, on a donc

$$\begin{aligned} D_T = -2L(\hat{\beta}) &= -2\left(N_{01}\log(\bar{\pi}_0) + N_{00}\log(1-\bar{\pi}_0) + N_{11}\log(\bar{\pi}_1)\right. \\ &\quad \left. + N_{10}\log(1-\bar{\pi}_1) + N_{21}\log(\bar{\pi}_2) + N_{20}\log(1-\bar{\pi}_2)\right) \end{aligned}$$

Le calcul donne $D_{T,obs} = 943.47$.

Pour le modèle vide, quelle que soit la valeur de $x \in \{0,1\}^2$, i.e. quelle que soit la classe de l'individu, on a $\mathbb{P}(Y = 1|X = x) = \exp(\beta_0)/(1 + \exp(\beta_0) = \pi_0$. Par conséquent, la vraisemblance du modèle vide est $V_\emptyset(\beta_0) = \pi_0^{N_{\bullet 1}}(1-\pi_0)^{N_{\bullet 0}}$, et la logvraisemblance est $L_\emptyset(\beta_0) = N_{\bullet 1}\log(\pi_0) + N_{\bullet 0}\log(1-\pi_0)$. Le maximum est atteint en $\tilde{\pi_0} = N_{\bullet 1}/n$ qui est la proportion empirique de malades dans l'échantillon. On obtient donc

$$D_\emptyset = -2\left(N_{\bullet 1}\log\left(\frac{N_{\bullet 1}}{n}\right) + N_{\bullet 0}\log\left(\frac{N_{\bullet 0}}{n}\right)\right).$$

Le calcul sur l'échantillon donne $D_{\emptyset,obs} = 1159$.

En supposant que les observations sont une réalisation du modèle total, on veut tester $(H_0) : \beta_1 = \beta_2 = 0$ contre $(H_1) : \beta_1 \neq 0$ ou $\beta_2 \neq 0$. L'hypothèse (H_0) revient à dire que les observations sont une réalisation du modèle vide, et que la probabilité de souffrir de problèmes de circulation sanguine ne dépend pas de la consommation de tabac. On fait le test du rapport de vraisemblance basé sur la statistique

$$TRV = 2L_T(\hat{\beta}) - 2L_\emptyset(\hat{\beta_0}) = D_\emptyset - D_T\,.$$

Sous (H_0) et pour n suffisamment grand, TRV suit approximativement une loi du χ^2 à 2 degrés de liberté, et la P-valeur du test est donc $P_{val} = \mathbb{P}(Z > TRV_{obs})$ où $Z \sim \chi^2_2$. $TRV_{obs} = D_{\emptyset,obs} - D_{T,obs} = 1159 - 943.47 = 215,53$, et $P_{val} = \mathbb{P}(Z > 215.53) \leq 5\%$. On conclut, en accord avec les résultats de la question 1., que la régression logistique est pertinente, et que les problèmes de circulation dépendent de la consommation journalière de tabac, avec 5% de chance de se tromper.

10. D'après les résultats généraux sur les estimateurs du maximum de vraisemblance, on sait que pour n assez grand, $\hat{\beta} \sim \mathcal{N}(\beta, (-D^2L(\hat{\beta}))^{-1})$, où $D^2L(\beta)$ est la matrice des dérivées secondes de L. Le calcul des dérivées secondes donne :

$$\begin{aligned}
-\frac{\partial^2 L}{\partial^2\beta_0} &= N_{0\bullet}\pi_0(1-\pi_0) + N_{1\bullet}\pi_1(1-\pi_1) + N_{2\bullet}\pi_2(1-\pi_2)\,;\\
-\frac{\partial^2 L}{\partial\beta_0\partial\beta_1} &= N_{1\bullet}\pi_1(1-\pi_1)\,;\\
-\frac{\partial^2 L}{\partial\beta_0\partial\beta_2} &= N_{2\bullet}\pi_2(1-\pi_2)\,;\\
-\frac{\partial^2 L}{\partial^2\beta_1} &= N_{1\bullet}\pi_1(1-\pi_1)\,;\\
-\frac{\partial^2 L}{\partial\beta_1\partial\beta_2} &= 0\,;\\
-\frac{\partial^2 L}{\partial^2\beta_2} &= N_{2\bullet}\pi_2(1-\pi_2)\,.
\end{aligned}$$

En notant

$$A := N_{0\bullet}\pi_0(1-\pi_0)\,,\ B := N_{1\bullet}\pi_1(1-\pi_1)\,,\ C := N_{2\bullet}\pi_2(1-\pi_2)\,,$$

on a donc

$$-D^2L(\beta) = \begin{pmatrix} A+B+C & B & C \\ B & B & 0 \\ C & 0 & C \end{pmatrix};$$

$$
\begin{aligned}
\det(-D^2L(\beta)) &= ABC\,;\\
(-D^2L(\beta))^{-1} &= \frac{1}{A}\begin{pmatrix} 1 & -1 & -1\\ -1 & 1+\frac{A}{B} & 1\\ -1 & 1 & 1+\frac{A}{C}\end{pmatrix}\\
&= \frac{1}{N_{0\bullet}\pi_0(1-\pi_0)}\begin{pmatrix} 1 & -1 & -1\\ -1 & \frac{N_{0\bullet}\pi_0(1-\pi_0)}{N_{1\bullet}\pi_1(1-\pi_1)}+1 & 1\\ -1 & 1 & \frac{N_{0\bullet}\pi_0(1-\pi_0)}{N_{2\bullet}\pi_2(1-\pi_2)}+1\end{pmatrix}.
\end{aligned}
$$

Pour $\beta=\hat\beta$, il vient

$$
\begin{aligned}
(-D^2L(\hat\beta))^{-1} &= \frac{1}{N_{0\bullet}\bar\pi_0(1-\bar\pi_0)}\begin{pmatrix} 1 & -1 & -1\\ -1 & \frac{N_{0\bullet}\bar\pi_0(1-\bar\pi_0)}{N_{1\bullet}\bar\pi_1(1-\bar\pi_1)}+1 & 1\\ -1 & 1 & \frac{N_{0\bullet}\bar\pi_0(1-\bar\pi_0)}{N_{2\bullet}\bar\pi_2(1-\bar\pi_2)}+1\end{pmatrix}.\\
&= \begin{pmatrix} \frac{N_{0\bullet}}{N_{01}N_{00}} & -\frac{N_{0\bullet}}{N_{01}N_{00}} & -\frac{N_{0\bullet}}{N_{01}N_{00}}\\ -\frac{N_{0\bullet}}{N_{01}N_{00}} & \frac{N_{0\bullet}}{N_{01}N_{00}}+\frac{N_{1\bullet}}{N_{11}N_{10}} & \frac{N_{0\bullet}}{N_{01}N_{00}}\\ -\frac{N_{0\bullet}}{N_{01}N_{00}} & \frac{N_{0\bullet}}{N_{01}N_{00}} & \frac{N_{0\bullet}}{N_{01}N_{00}}+\frac{N_{2\bullet}}{N_{21}N_{20}}\end{pmatrix}.
\end{aligned}
$$

Cela permet de tester pour $i\in\{0,1,2\}$, $(H_0):"\beta_i=0"$ contre $(H_1):"\beta_i\neq 0"$. Sous (H_0), $\hat\beta_i/\sigma(\hat\beta_i)$ suit approximativement une loi normale centrée réduite. De façon équivalente, sous (H_0), $\hat\beta_i^2/\sigma^2(\hat\beta_i)$ suit approximativement une loi du χ^2_1, et la Pvaleur du test est donnée par $P_{val}=\mathbb{P}\left[Z\geq\left(\hat\beta_i{}^2/\sigma^2(\hat\beta_i)\right)_{obs}\right]$, où $Z\sim\chi^2_1$.

Test de $(H_0):"\beta_0=0"$ contre $(H_1):"\beta_0\neq 0"$.
On a $\hat\beta_{0,obs}=-1.87$, $\sigma^2(\hat\beta_0)_{obs}=\frac{N_{0\bullet}}{N_{01}N_{00}}=\frac{300}{40(260)}=\frac{3}{104}$, soit $P_{val}=\mathbb{P}(Z\geq 104(1.87)^2/3)=\mathbb{P}(Z\geq 121.22)\leq 5\%$. On conclut donc que $\beta_0\neq 0$ avec 5% de chances d'avoir tort.

Test de $(H_0):"\beta_1=0"$ contre $(H_1):"\beta_1\neq 0"$.
On a $\hat\beta_{1,obs}=0.68$, $\sigma^2(\hat\beta_1)_{obs}=\frac{N_{0\bullet}}{N_{01}N_{00}}+\frac{N_{1\bullet}}{N_{11}N_{10}}=\frac{3}{104}+\frac{3}{161}=0.0474$, soit $P_{val}=\mathbb{P}(Z\geq(0.68)^2/0.0474)=\mathbb{P}(Z\geq 9.8)\leq 5\%$. On conclut donc que $\beta_1\neq 0$ avec 5% de chances d'avoir tort.

Test de $(H_0):"\beta_2=0"$ contre $(H_1):"\beta_2\neq 0"$.
On a $\hat\beta_{2,obs}=2.56$, $\sigma^2(\hat\beta_2)_{obs}=\frac{N_{0\bullet}}{N_{01}N_{00}}+\frac{N_{2\bullet}}{N_{21}N_{20}}=\frac{3}{104}+\frac{300}{100(200)}=\frac{3}{104}+\frac{3}{200}=0.0438$, soit $P_{val}=\mathbb{P}(Z\geq(2.56)^2/0.0438)=\mathbb{P}(Z\geq 150)\leq 5\%$. On conclut donc que $\beta_2\neq 0$ avec 5% de chances d'avoir tort.

Intervalle de confiance pour π_1.
Comme $\pi_1=F(\beta_0+\beta_1)$ et que F est strictement croissante, il suffit de construire un intervalle de confiance pour $\beta_0+\beta_1$, pour obtenir un intervalle de confiance pour π_1. Comme $\hat\beta$ est approximativement un vecteur gaussien, $\hat\beta_0+\hat\beta_1$ suit approximativement la loi gaussienne de moyenne $\beta_0+\beta_1$ et de variance $\sigma^2(\hat\beta_0+\hat\beta_1)=\sigma^2(\hat\beta_0)+\sigma^2(\hat\beta_1)+2\mathrm{cov}(\hat\beta_0,\hat\beta_1)=\frac{N_{1\bullet}}{N_{11}N_{10}}$. Autrement dit, pour n suffisamment grand $(\hat\beta_0+\hat\beta_1-(\beta_0+\beta_1))/\sqrt{N_{1\bullet}/N_{11}N_{10}}\sim\mathcal{N}(0,1)$. On peut parier avec 5% de chances de se tromper que

$$
\begin{aligned}
&-1.87+0.68-1.96\sqrt{\frac{300}{70(230)}}\leq\beta_0+\beta_1\leq-1.87+0.68+1.96\sqrt{\frac{300}{70(230)}}\\
\Leftrightarrow\quad &-1.457\leq\beta_0+\beta_1\leq-0.922
\end{aligned}
$$

$\Leftrightarrow \quad F(-1.457) \leq F(\beta_0 + \beta_1) = \pi_1 \leq F(-0.922)$

$\Leftrightarrow \quad 18.8\% \leq \pi_1 \leq 28.4\%$.

11. La table de confusion est la table de contingence des données $(y_i, \hat{y}_i)_{i=1,\cdots,900}$, où $\hat{y}_i$ est la valeur de y_i prédite par le modèle : $\hat{y}_i = \begin{cases} 1 & \text{si } \hat{\pi}(x_i) > 1/2\,, \\ 0 & \text{sinon.} \end{cases}$

Pour un individu non fumeur, $x_i = (0,0)$, $\hat{\pi}(x_i) = \bar{\pi}_0 = 2/15 \leq 1/2$. Le modèle prédit que les individus non fumeurs ne souffrent pas de problèmes de circulation.

Pour un individu fumant un paquet par jour, $x_i = (1,0)$, $\hat{\pi}(x_i) = \bar{\pi}_1 = 7/30 \leq 1/2$. Le modèle prédit que les individus fumant un paquet par jour, ne souffrent pas de problèmes de circulation.

Pour un individu fumant plus de deux paquets par jour, $x_i = (0,1)$, $\hat{\pi}(x_i) = \bar{\pi}_2 = 2/3 > 1/2$. Le modèle prédit que les individus fumant plus de deux paquets par jour, souffrent de problèmes de circulation.

On peut alors construire la table de confusion :

$y \setminus \hat{y}$	0	1
0	$N_{00} + N_{10} = 490$	$N_{20} = 100$
1	$N_{01} + N_{11} = 110$	$N_{21} = 200$

On en déduit le taux de vrais positifs $TVP = \frac{\sum_{i=1}^n \mathbb{1}_{y_i=1,\hat{y}_i=1}}{\sum_{i=1}^n \mathbb{1}_{y_i=1}} = \frac{200}{310}$, et le taux de faux positifs $TFP = \frac{\sum_{i=1}^n \mathbb{1}_{y_i=0,\hat{y}_i=1}}{\sum_{i=1}^n \mathbb{1}_{y_i=0}} = \frac{100}{590}$.

12. La courbe ROC du modèle est la courbe des points $(TFP(s), TVP(s))_{s\in[0,1]}$ quand on a adopté la règle d'affectation $\hat{y}_i = \begin{cases} 1 & \text{si } \hat{\pi}(x_i) > s\,, \\ 0 & \text{sinon.} \end{cases}$ Ici, il n'y a que trois valeurs possibles pour $\hat{\pi}(x_i)$, à savoir $\bar{\pi}_0 = 2/15$, $\bar{\pi}_1 = 7/30$, et $\bar{\pi}_2 = 2/3$.

Pour $s \in [0, 2/15[$, on a toujours $\hat{\pi}(x_i) > s$, et $\hat{y}_i = 1$ pour tous les individus de l'échantillon. On a donc $TVP(s) = \frac{N_{\bullet 1}}{N_{\bullet 1}} = 1$ et $TFP(s) = \frac{N_{\bullet 0}}{N_{\bullet 0}} = 1$.

Pour $s \in [2/15, 7/30[$, $\hat{\pi}(x_i) > s$ si et seulement si l'individu i est dans la classe C_1 ou C_2. On a donc $TVP(s) = \frac{N_{11}+N_{21}}{N_{\bullet 1}} = 270/310 = 0.87$ et $TFP(s) = \frac{N_{10}+N_{20}}{N_{\bullet 0}} = 330/590 = 0.56$.

Pour $s \in [7/30, 2/3[$, $\hat{\pi}(x_i) > s$ si et seulement si l'individu i est dans la classe C_2. On a donc $TVP(s) = \frac{N_{21}}{N_{\bullet 1}} = 200/310 = 0.645$ et $TFP(s) = \frac{N_{20}}{N_{\bullet 0}} = 100/590 = 0.17$.

Pour $s \in [2/3, 1]$, on a toujours $\hat{\pi}(x_i) \leq s$, et $\hat{y}_i = 0$ pour tous les individus de l'échantillon. On a donc $TVP(s) = TFP(s) = 0$.

La courbe ROC est donc constituée des quatre points (0,0), (0.17,0.645), (0.56,0.87), (1,1). Elle est représentée sur la Figure 1.

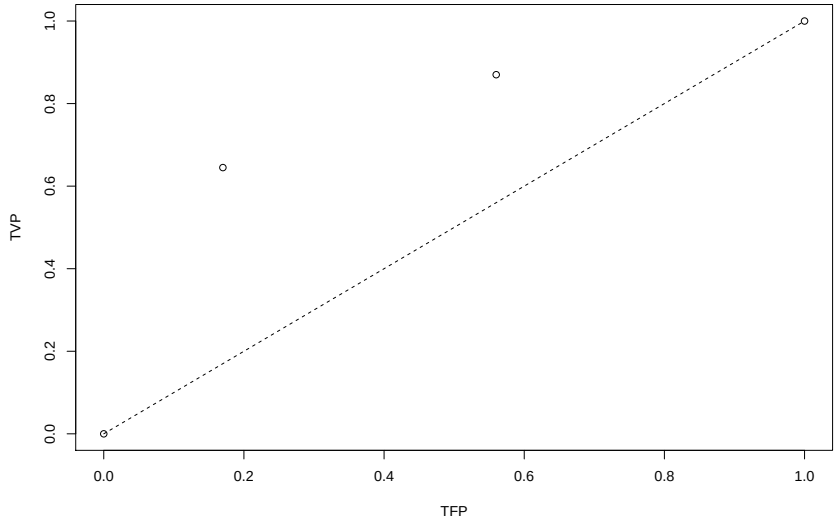

FIGURE 1 – Courbe ROC du modèle complet.